**Computer Stylometry of C.S. Lewis's *The Dark Tower* and Related Texts.**

**Michael Oakes, RIILP, University of Wolverhampton, England.**

**Abstract**

This paper looks at the provenance of the unfinished novel *The Dark Tower*, generally attributed to C.S. Lewis. The manuscript was purportedly rescued from a bonfire shortly after Lewis's death by his literary executor Walter Hooper, but the quality of the text is hardly vintage Lewis. Using computer stylometric programs made available by Eder et al.'s (2016) "stylo" package and a word length analysis, samples of each chapter of *The Dark Tower* were compared with works known to be by Lewis, two books by Hooper and a hoax letter concerning the bonfire by Anthony Marchington. Initial experiments found that the first six chapters of *The Dark Tower* were stylometrically consistent with Lewis's known works, but the incomplete chapter 7 was not. This may have been due to an abrupt change in genre, from narrative to pseudoscientific style. Using principal components analysis, it was found that the first and subsequent components were able to separate genre and individual style, and thus a plot of the second against the third principal components enabled the effects of genre to be filtered out. This showed that chapter 7 was also consistent with the other samples of C.S. Lewis's writing.

**Introduction**

Clive Staples Lewis (1898 – 1963) was a prolific writer, and his best-loved fiction is probably his Deep Space trilogy, *The Screwtape Letters*, and his "Narnia" series of children's books. Shortly after Lewis' death, Walter Hooper, the literary executor for the Lewis Estate, claimed to have found an unpublished fragment of fiction, which was published much later (1977) as *The Dark Tower*. There is some overlap between *The Dark Tower* and the Deep Space trilogy, as they share a number of characters such as McPhee, Ransom and even Lewis himself.

For many years, C. S. Lewis had lived with his brother Warren at a house called the Kilns, in Oxford. In the first paragraph of the preface to the version of *The Dark Tower* published by Fount, Hooper claimed that Warren wanted to dispose of his late brother's old papers, and ordered the gardener to light a bonfire of them which "burned steadily for three days". In Hooper's own words,

"Happily, however, the Lewis's gardener, Fred Paxford, knew that I had the highest regard for anything in the master's hand, and when he was given a great quantity of CS Lewis's notebooks and papers to lay on the flames, he urged the Major [Warren Lewis] to delay till I should have a chance to see them" One of the rescued notebooks contained the hand-written manuscript of *The Dark Tower*" (Hooper, 1977:vii).

*The Dark Tower* was eventually published with a number of C.S.Lewis short stories, all of which had been published before, except for the very brief *The Man Born Blind* which had been found in a notebook given to Walter Hooper by Lewis's brother.

The story of the bonfire was later denied by Fred Paxford, and this denial was published in the journal *Christianity and Literature* (Lindskoog, 1978). Shortly afterwards, *Christianity and Literature* (1979) also published a letter from Anthony Marchington, seemingly in support of Paxford's denial, as it stated that a chemical analysis of the soil in Lewis' garden had revealed that no major bonfire

had been lit there. This letter is thought to be a hoax: its content is clearly pseudo-scientific, and Marchington was a close friend of Walter Hooper, at one time sharing lodgings with him. *The Dark Tower* itself is unfinished, possibly because the plot hits something of a dead end. Opinions vary as to the quality of the writing, and the story changes tack abruptly in the final chapter, where the protagonist Scudamour is left alone in a library in "Othertime" to learn about the "Othertimers'" discoveries about time travel. Hooper (1977: viii) estimates that Lewis began writing *The Dark Tower* soon after completing *Out of the Silent Planet* in 1938. There are similarities with Madeleine L'Engle's *A Wrinkle in Time*, although this was not written until 1962. All this has led a number of people, most notably Katherine Lindskoog, to conclude that *The Dark Tower* may not be entirely written by C.S. Lewis. The most likely candidates for writing at least parts of *The Dark Tower*, apart from Lewis himself, would be Walter Hooper and Anthony Marchington. Lindskoog (1988:53-54) mainly suspects Marchington:

"No one thinks that Walter Hooper could have tackled all that ficto-science. The most obvious suspect is Anthony Marchington himself. He is a scientist, he is interested in the origin of *The Dark Tower*, and he has tricked *Christianity and Literature* with a scientific spoof. Furthermore, he was about eight years old when Madeleine L'Engle published her children's classic *A Wrinkle in Time*, and so he quite possibly read it as a child. That could account for unconscious copying of Engle's automaton scene in *The Dark Tower*." The corresponding "automaton scene" in The Dark Tower occurs in Chapter 2.

**Previous Work**

In the past, a number of computer stylometric analyses have been performed on *The Dark Tower* and related texts. The first of these was by Carla Faust Jones (1989), who used a computer program written by Jim Tankard which he had previously used to study the Federalist Papers (Tankard, 1986). First the program finds the frequencies of character n-grams (sequences of n consecutive characters, where n was 1 or 2) in the text, then normalises these to frequencies per 1000 characters, rounded to nearest whole number. Spaces and punctuation were not considered, and upper and lower case characters were considered equivalent. For 1-grams (single characters), the index of difference between two text samples was given by the expression

$$\sum_{a}^{z} |f_A - f_B|$$

Where $f_A$ is the frequency of a character in the first text sample, and $f_B$ is the frequency of that character in the second. The differences in these frequencies are found for every character in the alphabet, then all added together. For the 2-grams the expression is analogous: we find the differences in the frequencies of every possible character pair in the two texts, then add together all 26 x 26 differences. Jones' (1989) results are shown in Table 1.

| Comparison | Texts Compared | D. I. (unigrams) | D. I. (bigrams) |
|---|---|---|---|
| A1 | Silent Planet & Perelandra | 76 | 1778 |
| A2 | Silent Planet & | 60 | 1890 |

| | | | |
|---|---|---|---|
| | Hideous Strength | | |
| A3 | Perelandra & Hideous Strength | 74 | 1834 |
| B1 | Silent Planet & Dark Tower | 113 | 2427 |
| B2 | Perelandra & Dark Tower | 83 | 2137 |
| B3 | Hideous Strength & Dark Tower | 91 | 2327 |

Table 1. Indexes of difference between *The Dark Tower* and C.S.Lewis's three complete science fiction novels, found by Jones (1989).

Both the 1-gram and 2-gram analyses show that the three complete science fiction novels by Lewis, *Out of the Silent Planet*, *Perelandra* and *That Hideous Strength* are more similar to each other than they are to *The Dark Tower*. Although this is interesting, it does not prove that *The Dark Tower* was not written by Lewis. There is no comparison with Lewis's other works, nor any comparison with works by other candidate authors for *The Dark Tower*.

Lindskoog (1994:247-8) describes a seemingly unpublished report by Andrew Queen Morton. He used a data visualisation technique called a cusum analysis, which has been used to detect changes in the quality of production line outputs in an industrial setting. Morton himself suggested that this technique could be used to detect discontinuities in writing style, such as when one author breaks off and another begins in a multiple-authored text. A linguistic feature such as word length or noun frequency is used to characterise the texts. The resulting graph shows an upward trend for those portions of the text which show an above average (taken over the text as a whole) occurrence of the chosen feature, and a downward trend for those parts which show a below average occurrence. Thus, if two authors who have contributed to a text show different rates of usage of the chosen feature, the point where one writer hands over to another might show an abrupt change in the direction (upwards or downwards) of the graph.

Morton took the first 23 sentences of chapter one of *The Dark Tower*, the first 24 sentences of chapter four, and the first 25 sentences of chapter seven, alongside sections from *Out of the Silent Planet* and *That Hideous Strength*. Morton concluded that *The Dark Tower* was a composite work: Lewis did not write chapters one and four, but he did write chapter seven, the one with the library scene. The technique is highly controversial in studies of disputed authorship, but my feeling is that the choice of linguistic features may affect the success of the technique itself. For example, Merriam (2000) achieved interesting results for the Shakespeare play Edward III with cusum charts using the frequencies of prosodic features, rare words and function words, combined into a single chart using Principal Component Analysis (PCA). Unfortunately Lindskoog gives no details of which linguistic features Morton used to characterise the texts. Morton's study also suffers from the brevity of the texts which were analysed.

More recently, Thompson and Rasp (2009) used statistical techniques developed by Thisted and Efron (1987) for comparing smaller samples of unknown authorship (such as a newly discovered

text) with a much larger canon with known authorship. If we define *t* as the size in words of the small sample divided by the size in words of the larger canon, $n_1$ as the number of words occurring exactly once in the canon, $n_2$ the number of words occurring twice, and so on, then in their "new words" test, we can estimate $\widehat{v_0}$, the number of words in the smaller text that do not appear in the larger canon, as follows:

$$\widehat{v_0} = n_1 t - n_2 t^2 + n_3 t^3 \ldots$$

This formula depends on t being small, to ensure that the series converges. We want to see how close the estimated value of $\widehat{v_0}$ is to $m_0$, which is the number of "new" words actually found in the small sample but not in the canon. If these values differ greatly, it suggests that the small sample was not written by the author of the canon.

They performed three other tests using related formulae – the "rare words" test, where the estimated and true numbers of words occurring below an arbitrary threshold number of times are compared; and the "slope" and "uniformity" tests, which take into account the estimated and real numbers of words of every individual frequency up to a threshold. The tests were validated first by comparing samples of George MacDonald's writings with those known to be by Lewis. The "new words" test was most successful, being able to discriminate between them 25% of the time with 95% confidence – we would expect only 5% if we were looking at a single author. The "new words" test also showed the best discriminatory power between short samples of *The Dark Tower* and two of Lewis' science fiction novels (those thought to have been written closest in time to *The Dark Tower*), namely *Out of the Silent Planet* and *Perelandra*. The test found that 29% of the *Dark Tower* samples were significantly different to the "canon" of two science fiction novels. Overall, Thompson and Rasp felt that their results were inconclusive. Even though the "new words" test did discriminate between samples of *The Dark Tower* and the complete novels, this may not have been due to a difference in authorship, but because a novel in draft form might differ from a complete, polished work.

**Stylometry with R: the 'Stylo' Package**

Before describing the specific experiments carried out for this paper, I will describe some general features of the package that were used, "Stylometry with R" (stylo), which was written in the R statistical programming language by Eder et al. (2016). Stylo enables a choice of measures of document dissimilarity, and I used the classic Burrows' Delta, first described by Burrows (2002), throughout. Stylo also allows a variety of linguistic features to be used to characterise the texts, these being word and character overlapping n-grams, where n can be any number, including 1 for single words or characters. An n-gram is a sequence of n tokens. For example, if n is 2 and we are interested in overlapping character sequences, a word like "Lewis" would be analysed into the four entities "Le", "ew", "wi" and "is". Finally, stylo enables a number of kinds of graphical displays, each of which is a way of showing which documents are most similar to each other, by placing them close together on the page. For example, Figure 1 is an example showing the outputs for hierarchical agglomerative clustering (HAC). The relationships between the texts are shown on dendrograms, so called because they look like trees on their side. The branches on the extreme right each correspond to individual texts, and texts on nearby branches are similar to each other. The technique for building a dendrogram is to first find the most similar pair of texts, and join them together, so that thereafter they can be considered as a joint entity. In the subsequent series of steps, each time the most similar pair of single texts or joint entities is fused to form a larger group. This process

continues until all the texts are joined in a single structure. When using Ward's (1963) method, the default linkage method offered by stylo, the document similarities between a newly formed joint entity and all the other text groups formed so far is a function of the distances between each of the two constituents before fusion and the rest of the text groupings, and the number of texts in each entity. A series of dendrograms obtained for different numbers of linguistic features can be fused into a bootstrap consensus tree, such as that shown in Figure 2. Branches between texts are shown whenever such a branch was found in a selected proportion of the dendrograms – I used the default value of 0.5 throughout. A third type of representation, called Principal Components Analysis, can be seen for example in Figure 3. The technique aims to find groups of texts which are characterised by the common presence or absence of certain groups of linguistic features, which form a component. Texts with many of these features score highly on the component, while other texts with few of them have negative scores on this component. The component which accounts for the greatest amount of variability between the texts is called the first principal component (PC1), but there are other components which successively account for less variability between the texts. Normally the texts are plotted according to their positions on the first two components (PC1 and PC2), but as we shall see in this paper, if PC1 corresponds to genre rather than author, genre effects can sometimes be overcome by plotting the texts according to their scores on lower components (such as PC2 and PC3). PCA is often used to examine variation in language. For example, Holmes et al. (2002) used PCA to examine authorship of the "Pickett letters" from the American Civil War, Binongo and Smith (1999) used PCA to study the authorship of the play *Pericles*, and Harris (2010) looked at possible genres in the corpus of Rongorongo from the Easter Islands. Biber (1988) used the closely related technique of Factor Analysis to study functional linguistic variation arising from genre and register. Stylo allows a culling parameter to be set. For example, if this value is 20, then only features appearing in at least 20% of the texts will be considered in the analysis. In all the experiments described in this paper, the "culling" parameter was set to 0, so for example if we are studying the frequencies of the top 100 words, the frequency of every one of these words will be considered. The 100 most frequent words are the 100 most frequent words in the entire corpus, rather than the 100 most frequent words in an individual sample. It is possible to use text samples of different sizes, because the word frequencies are normalised.

Throughout the experiments the following text pre-processing steps were adhered to. By selecting the "English" button on the "Input and Language" page of the stylo Graphical User Interface (GUI), contractions such as "don't" will be treated as the two single words "don" and "t". Hyphenated compound words such as "topsy-turvy" also become two single words, here "topsy" and "turvy" (Eder et al., 2015:11). The "Preserve Case" button was not selected, so all upper case characters were converted to lower case. I did not select the option to delete pronouns, and no stop list was used, but did select the option to read in text as plain text files. By default, all sequences of non-alphabetic characters were reduced to a single white space for n-grams longer than 1. Single words were treated as single letters separated by spaces. It is possible to examine the full feature set with the R command `stylo.results = stylo()` , then running the GUI to select the desired feature set, and then examining the set with `stylo.results$features` (Eder et al., 2016: 112).

**Text Samples**

| Samples | Author | Year | Title | Sample Length |
|---|---|---|---|---|

| | | | | (words each) |
|---|---|---|---|---|
| LOR0, LOR1, LOR2, LOR3 | J.R.R. Tolkein | 1954-1955 | Lord of the Rings | 7381, 9820, 3375, 13039 |
| HOB1, HOB2, HOB3, HOB4 | J.R.R. Tolkein | 1937 | The Hobbit | 8652, 5234, 2874, 4066 |
| ENG1, ENG2, ENG3, ENG4 | Madeleine L'Engle | 1962 | A Wrinkle in Time | 4652, 3628, 3819, 4198 |
| LWW, DAWN, MN, LB | C. S. Lewis | 1950-1956 | "Narnia" Series | 3869, 3237, 3035, 2648 |
| THS1, THS2, THS3, THS4 | C. S. Lewis | 1945 | That Hideous Strength | 9069, 7468, 8941, 8457 |
| PER1, PER2, PER3, PER4 | C. S. Lewis | 1943 | Perelandra | 4980, 4208, 5329, 5508 |
| OSP1, OSP2, OSP3, OSP4 | C. S. Lewis | 1938 | Out of the Silent Planet | 5497, 3635, 3761, 3794 |
| MBB | C. S. Lewis | Unknown | The Man Born Blind | 1769 |
| LEFAY | C. S. Lewis | Unknown | The "Lefay" fragment | 5437 |
| DT1, DT2, DT3, DT4, DT5, DT6, DT7 | C. S. Lewis | Unknown | The Dark Tower | 3010, 4190, 4879, 5645, 3504, 3580, 3691 |
| TJB1, TJB2, TJB3, TJB4 | Walter Hooper | 1982 | Through Joy and Beyond | 4532, 6676, 3918, 5400 |
| PWD1_2, PWD3_4, PWD5, PWD6 | Walter Hooper | 1971 | Past Watchful Dragons | 4574, 3991, 2299, 4689 |
| MLET | Tony Marchington | 1979 | Letter to "Christianity and Literature" | 986 |
| MC1, MC2 | C. S. Lewis | 1942-1944 | Mere Christianity | 8229, 9192 |
| PP1, PP2 | C. S. Lewis | 1940 | The Problem of Pain | 3677, 3001 |

Table 2. Text samples used in the experiments described in this paper.

The set of text samples used in these experiments are summarised in Table 2. The four texts from *Lord of the Rings* are the Prologue, and the first chapter of each of three parts (called individually *The Fellowship of the Ring, The Two Towers* and *The Return of the King*). The texts from the *The Hobbit* are chapters 1 to 4, and the texts from the Narnia series are the first two chapters of *The Lion, The Witch and the Wardrobe*, the first chapter of *The Voyage of the Dawn Treader*, the first chapter of *The Magician's Nephew*, and the first chapter of *The Last Battle*. The four samples of *That Hideous Strength* are the first four chapters , as is the case for *Perelandra.* However, the four samples of *Out of the Silent Planet* consist of the first two chapters; the third and fourth chapters; the fifth and sixth chapters; and the seventh and eighth chapter. The two shortest texts, *The Man Born Blind* and the Marchington letter were used in their entirety, as was the *Lefay* fragment. The seven samples of *The Dark Tower* consist of one chapter each, including the seventh and final (but unfinished) chapter. The four samples of *Through Joy and Beyond* consist of one part each of that

book – and thus comprise the entirety of that book. The four samples of *Past Watchful Dragons* consist of: the first two chapters; the third and fourth chapters; the fifth chapter, excluding the *Lefay* fragment; and the sixth chapter. The two samples of *Mere Christianity* are books 1 and 2 (*Right and Wrong as a Clue to the Meaning of the Universe* and *What Christians Believe*). Finally, the two samples of *The Problem of Pain* are chapters one and two of that book.

The Lewis texts are compared against the Tolkein texts because the two authors were close friends who regularly discussed their work at meetings of the literary group called "The Inklings", which met at the "Eagle and Child" pub in Oxford. They both wrote about other worlds, such as Middle Earth (Tolkein) and Narnia (Lewis). Like Lewis and Tolkein, Madeleine L'Engle also wrote children's fantasy novel with a Christian theme, where children are transported to faraway planets. As stated in the introduction, Lindskoog has noticed similarities between *The Dark Tower* and *A Wrinkle in Time*. The Lefay fragment is a long fragment of a draft of the sixth Narnia book, *The Magician's Nephew*, also found by Hooper in one of Lewis's notebooks, and reproduced in *Past Watchful Dragons* (Hooper, 1971: 48-65). *Out of the Silent Planet*, *Perelandra* and *That Hideous Strength* are Lewis's three science fiction works for adults. As described above, Walter Hooper claimed to have discovered the short story *A Man Born Blind* and the unfinished novel *The Dark Tower* in notebooks, written in Lewis's handwriting, after Lewis's death. However, the handwriting in this notebook has never been satisfactorily authenticated (Lindskoog, 1999). The two selected works by Walter Hooper himself are *Past Watchful Dragons*, a guide to the Narnia books, and *Through Joy and Beyond*, a biography of C.S. Lewis. A further sample used is the full text of Tony Marchington's hoax letter to *Christianity and Literature.* To the author's best knowledge, Tony Marchington left no other published works, and thus it was not possible to use a larger sample of Marchington's writing in these experiments. *Mere Christianity* and *The Problem of Pain* are examples of Lewis's non-fiction writing.

**Experiments 1: Discrimination between Lewis and Two Other Authors of Fiction.**

The first set of experiments, the baseline, were designed to show whether the multivariate statistical techniques available in the stylo package were able to distinguish between the three authors L'Engle, Tolkien and Lewis. The results for the hierarchical clustering (Ward's method) using the 100 most frequent words (MFW) as linguistic features are shown in Figure 1. The choice of 100 words is made following the recommendations of Burrows (2002) and Juola (2015: i108), as the 100 most frequent words are typically function words, giving information about grammar and individual writing style rather than content.

{FIGURE 1 ABOUT HERE}

Here we see four main clusters, which from top to bottom correspond to a) Tolkein, with the exception of Lewis's *Voyage of the Dawn Treader*; b) Lewis's first two books from the Deep Space Trilogy; c) children's books written by L'Engle and Lewis, except for the second section of *Perelandra*; and d) Lewis's last book from the Deep Space Trilogy. The same pattern is seen more clearly in the bootstrap consensus tree (also for the 100 most frequent single words), as shown in Figure 2, where the two "Deep Space" branches are placed closer together, effectively leaving three main clusters in the diagram.

{FIGURE 2 ABOUT HERE}

In Figure 3, the data for 100 most frequent single words is displayed using Principal Components Analysis. Once again we see three main groupings, with samples by Lewis seen in the top right of the diagram, samples Children's books in the middle left part, and samples by Tolkein in the bottom right section.

{FIGURE 3 ABOUT HERE}

Juola (2015) recommends running a series of independent analyses in stylometric work. While a series of runs using the same feature set with different clustering algorithms (such as shown in Figures 1 to 3) are not independent of each other, experiments using distinct feature sets would be. In his experiment on the writing of J.K. Rowling, Juola states that "Tests were run on four separate feature sets: word lengths, character 4-grams, word pairs, and the 100 most frequent words" (Juola, 2015:i108). Juola (personal communication) recommends using all character 4-grams and word pairs, not just the top *n*. To achieve this as far as possible, I set *n* to the very high value of 5000. Although these linguistic features are not completely independent of each other (for example if a word has high frequency, this will raise the frequencies of its constituent n-grams), I endeavoured to follow his approach. The groupings produced by the hierarchical clustering when using either the 5000 most frequent word 2-grams (see Figure 4) or the 5000 most frequent character 4-grams (shown in Figure 5) were the same as each other, producing somewhat clearer separation between the authors than was the case for the 100 most frequent single words.

{FIGURE 4 ABOUT HERE}

{FIGURE 5 ABOUT HERE}

In Figures 4 and 5 we again see a cluster for children's authors, but this time we see more separation between those in Lewis's Narnia series and those by Madeleine L'Engle, than we saw in Figures 1 to 3. The middle cluster consists entirely of Tolkein samples, and the bottom cluster contains all the books in Lewis's Deep Space Trilogy. Thus it seems that it is possible to some extent to distinguish between the three authors of fiction, but the situation is partly confused because we are seeing both the effects of authorship and of genre. As a result we have two clusters for Lewis, one for his adult fiction, and another for his children's fiction, which is only marginally distinguished from another author (L'Engle) who also wrote in the children's fiction genre. To separate authorship and genre, it is possible to use the technique of principal component analysis (PCA). For example, Schöch used PCA to examine French plays by the brothers Pierre and Thomas Corneille. The first principal component separated the plays by author, but the second component separated them by genre, tragedy or comedy. An example of a feature which distinguished the plays by genre was the word *mort* (death) which was much more prevalent in tragedies than comedies. One of the features discriminating between the two authors was the function word *ces* (these). Using the related technique of Correspondence Analysis, Linmans (1998) showed that samples taken from the Synoptic Gospels were separated on the first component according to genre (discourse, aphorisms, narrative or parable), and on the second component according to author (Mathew, Mark or Luke). I ran a PCA on the most frequent 5000 character 4-grams data, and achieved the plot shown in Figure 6.

{FIGURE 6 ABOUT HERE}

As in all the previous experiments, we see three main groupings in the text samples. This time the Tolkein samples all appear in the top half of the plot, the children's writing appears in bottom left part, and the Deep Space samples by Lewis appear in the bottom right part. Thus Lewis's texts still appear in two separate clusters – one for his children's writing, and one for his adult science fiction. At the most coarse grained division of texts, we see writing for children in the left half of the diagram, corresponding to negative scores on PC1, and writing for adults in the right division, corresponding to positive scores on PC1. Although *Lord of the Rings* was not specifically written for children, it was written as a sequel to *The Hobbit*, which was. Thus we see the samples of *The Hobbit* appearing to the left of those from *Lord of the Rings*. In this experiment discrimination by genre was seen to be more pronounced than discrimination by author, since the first principal component (PC1) accounts for more variation in the data than any of the other principal components. We can remove the effect of genre by taking PC1 out of the diagram, and instead of plotting PC1 against PC2, plotting PC2 against PC3, as shown in Figure 7. There is no option on the stylo GUI for plotting PCA components other than the first and second, but this may be done with the following series of R commands:

```
> a = stylo()
> b = a$pca.coordinates
> PC2 = b[,2]
> PC3 = b[,3]
> labels = names(PC2)
> plot(PC2, PC3, pch=" ")
> for (i in 1:length(labels)){
+ text(PC2[i], PC3[i], labels[i])
+ }
```

This has the effect of grouping all the Lewis texts together, irrespective of genre, in the bottom left of the diagram. There is now also a distinct cluster for L'Engle in the top left corner, and the Tolkein samples all appear on the right hand side.

{FIGURE 7 ABOUT HERE}

**Experiments 2: The Dark Tower in Relation to Texts by Lewis, Hooper and Marchington.**

The second set of experiments was designed to show where the individual chapters of *The Dark Tower* lay in relation to known works by Lewis, Hooper and the Marchington letter. The results are shown in Figure 8 for hierarchical clustering by Ward's method with 100 most frequent single words. The coarsest (leftmost) subdivision separates most of the known works by Lewis from those by Hooper and Marchington. The posthumously discovered texts (MBB, LEFAY and DT1 to DT6) all cluster very close together, and all are well within the main Lewis cluster. This suggests that all these texts were indeed written by Lewis. The main surprise was that there was a small cluster of Lewis texts at the bottom, attached to the Hooper/Marchington cluster. The final chapter of *The Dark Tower* (DT7) appeared in this small cluster, and thus seems to have stylistic similarities with works both by and not by Lewis. The experiment was repeated using the 5000 most frequent character 4-grams, since this feature gave the most clear cut results for the fiction texts. These results are shown in Figure 9.

{FIGURE 8 ABOUT HERE}

{FIGURE 9 ABOUT HERE}

The results are more clear cut when using the 5000 most frequent character 4-grams (Figure 9) than when using the top 100 single words (Figure 8), and give two main clusters. All samples of Lewis's known fiction appear in the bottom cluster, along with the posthumously published samples MBB, LEFAY and chapters 1 to 6 of *The Dark Tower*. The top cluster contains all the samples of Hooper's works, clustered tightly together, the Marchington Letter, and a tight grouping containing Chapter 7 of *The Dark Tower* and 4 samples of Lewis's non-fiction. The main division between the texts thus appears to be non-fictional (top cluster) versus fictional (bottom cluster). Once again we have a situation where genre and authorship confound each other – does the final chapter of The Dark Tower appear in the top cluster because it is written by Hooper or Marchington, or because it is written in the style of non-fiction? The next step was to perform PCA experiments to first try and determine whether the first principal component (PC1) did indeed correspond to genre, and if so omit this component from a future analysis using PC2 and PC3. This would ideally extract the effects of genre, so that the results of authorship alone can be seen. The PCA analysis plotting the text samples according to their scores on the second and third principal components are shown in Figure 10. The fictional text samples have all got positive (or only slightly negative) scores on PC1, and the non-fiction samples almost all negative (or only slightly positive) scores on PC1. Thus PC1 is polarised by genre, and was eliminated at the next step. DT7 very close to works by Hooper, being almost superimposed on the cluster of text samples by Hooper in the bottom left quadrant. Is this because they were actually written by Hooper, or are they simply written in a stylistically similar non-fictional style? The small Marchington sample appears as a complete outlier at highly negatives scores on both PC1 and PC2.

{FIGURE 10 ABOUT HERE}

In the next experiment, I removed the effect of genre which gave the polarity seen on PC1, where all the non-fiction texts are placed on the left hand side, and all the fiction texts are placed on the right hand side. This was done by omitting PC1, and plotting PC2 against PC3. This plot is shown in Figure 11. This plot was inconclusive, since the Hooper and Lewis samples appeared very close together (albeit with a tendency for the Hooper samples to appear near the top), and DT7 is almost equidistant between samples by the two authors Hooper and Lewis. Further experimentation showed that Figure 11 was probably distorted due to the outlying Marchington letter (M_LET) sample, which was much smaller than the others and thus probably contained much statistical noise. In addition, while it was pseudo-scientific in style it was also a letter, which would also put it in contrast with the other texts. After removing this sample, the character 4-gram frequencies in the corpus were recalculated to include only the remaining texts. When this sample was removed, I obtained the much clearer plot shown in Figure 12. Here the Hooper texts are plotted at positive values of both PC2 and PC3, and thus form a cluster in the top right part of the graph. DT7 now plots much closer to the Lewis texts.

{FIGURE 11 ABOUT HERE}

{FIGURE 12 ABOUT HERE}

Figure 12. Plot of The Dark Tower Chapters and Texts by Lewis and Hooper on the Second and Third Principal Components, Using the 5000 Most Frequent Character 4-Grams.

**Experiments 3: Word Length Experiments**

The one linguistic feature suggested by Juola (2015:i108) not yet examined in this paper is mean word length. The mean word lengths (in characters) for each of the text samples used in this paper were found, using a program written in PERL by the author (see ANONYMISED WEB PAGE), and are shown in Table 3.

| Text Sample | Words | Characters | Average word length |
|---|---|---|---|
| M_LET | 986 | 4775 | 4.843 |
| DT7 | 3691 | 16696 | 4.523 |
| PWD6 | 4689 | 21110 | 4.502 |
| PWD1_2 | 4574 | 20518 | 4.489 |
| PWD3_4 | 3991 | 17844 | 4.471 |
| PP1 | 3677 | 16422 | 4.466 |
| PP2 | 3001 | 13317 | 4.378 |
| OSP3 | 3761 | 16680 | 4.345 |
| TJB1 | 4532 | 20075 | 4.430 |
| TJB4 | 5400 | 23894 | 4.425 |
| TJB3 | 3918 | 17336 | 4.425 |
| TJB2 | 6676 | 29457 | 4.412 |
| PWD5 | 2299 | 10098 | 3.392 |
| LOR0 | 7381 | 32396 | 4.389 |
| OSP4 | 3794 | 16612 | 4.378 |
| THS1 | 9069 | 39528 | 4.359 |
| THS2 | 7468 | 32433 | 4.343 |
| PER4 | 5508 | 23906 | 4.340 |
| ENG4 | 4198 | 18028 | 4.294 |
| PER3 | 5329 | 22866 | 4.291 |
| OSP2 | 3635 | 15559 | 4.280 |
| DT1 | 3010 | 12837 | 4.265 |
| HOB4 | 4066 | 17287 | 4.252 |
| LOR1 | 9820 | 41593 | 4.236 |
| THS4 | 8457 | 35807 | 4.234 |
| DT3 | 4879 | 20614 | 4.225 |
| THS3 | 8941 | 37728 | 4.220 |
| DT6 | 3580 | 15105 | 4.219 |
| DT2 | 4190 | 17670 | 4.217 |
| ENG1 | 4652 | 19518 | 4.196 |
| PER1 | 4980 | 20862 | 4.189 |
| ENG2 | 3628 | 15171 | 4.182 |
| OSP1 | 5497 | 22940 | 4.173 |
| DAWN | 3237 | 13443 | 4.153 |
| DT4 | 5645 | 23388 | 4.143 |
| ENG3 | 3819 | 15799 | 4.137 |
| HOB1 | 8652 | 35654 | 4.121 |
| HOB2 | 4234 | 21436 | 4.096 |
| HOB3 | 2874 | 11761 | 4.092 |
| DT5 | 3504 | 14327 | 4.089 |
| MC2 | 9192 | 37505 | 4.080 |
| LOR2 | 3375 | 13685 | 4.055 |

| | | | |
|---|---|---|---|
| LOR3 | 13039 | 52740 | 4.045 |
| MBB | 1769 | 7125 | 4.028 |
| MC1 | 8229 | 33106 | 4.023 |
| PER2 | 4208 | 16905 | 4.017 |
| MN | 3035 | 12112 | 3.991 |
| LEFAY | 5437 | 21676 | 3.987 |
| LWW | 3869 | 15258 | 3.944 |
| LB | 2648 | 10395 | 3.926 |

Table 3. Average word lengths (in characters) for each of the text samples

Although average word length is often considered a blunt tool for assigning authorship, the results in Table 3 generally accord with the experiments performed on stylo, before the effects of genre were filtered out by PCA. The 13 texts with greatest average word length are all non-fictional, which is the style in which DT7 is written. The texts by Hooper and Marchington, and the final chapter of *The Dark Tower* are grouped together at the top of the table as they have greater average word length than the other texts. The texts with lowest average word length are the Narnia series, including the Lefay fragment. The other children's authors also tended to use shorter words: the average word lengths for the Madeleine L'Engle samples were in the range 4.134 to 4.294; for Tolkein's *The Hobbit* the range was 4.092 to 4.252; and for Tolkein's *Lord of the Rings* it was from 4.045 to 4.294, except for the Prologue which was 4.389. Since word length is a single figure which depends on both genre and authorship, it is not possible to separate these out using this technique alone, and thus the final chapter of *The Dark Tower* appears close to the Marchington and Hooper samples, possibly because they are all written in the style of non-fiction. To filter out the effect of genre, it might be possible to find the mean word lengths for the genres (children's fiction, adult fiction, adult non-fiction) over a large range of authors, and to find the word lengths of our samples relative to these means. Word length as a feature has been found in several multi-dimensional studies, such as Biber (1988), revealing that word length has functional properties.

**Conclusion**

From these analyses, I feel that it is clear that Lewis wrote the first 6 chapters of *The Dark Tower*, as well as *The Man Born Blind* and the Lefay fragment, all of which were found by Walter Hooper in notebooks after Lewis's death. Initial results did show that the final chapter of *The Dark Tower* was more stylistically consistent with the samples of Hooper and Marchington's writing. However, this may be more a question of genre than authorship, since the plot of *The Dark Tower* changes abruptly from a narrative account in the first six chapters, to a pseudoscientific description of how the people of "Othertime" discovered time travel in the seventh chapter. Marchington's letter is also in pseudoscientific style, as it describes the results of a (ficticious) soil analysis. Although Hooper's texts are not pseudoscientific, they are not narrative fiction either, which may explain why they initially clustered with the Marchington letter and the final chapter of *The Dark Tower*. The use of PCA where factors corresponding to genre were not plotted proved to be an effective means of filtering out genre. Once the effects of genre were removed, text sample DT7 did appear to be more typical of the Lewis texts than the Hooper texts. Discovering the contents of a library in another world is in fact a Lewisian motif, seen in *The Voyage of the Dawn Treader*, the third of the Narnia series, when Lucy reads the contents of a book of magical spells in the library of the fallen star

Coriakin. On the other hand, if an unfinished work were to be added to, it would be easier to add a new chapter at the end than at any other place in the text.

**References**


Biber, D. (1988). *Variation across Speech and Writing*. Cambridge, UK: Combridge University Press.

Binongo, J. and Smith, M.W.A. (1999). The application of principal component analysis to stylometry. Literary and Linguistic Computing 14(4):445-465.

Burrows, J. (2002). Delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing* **17**(3): 267-287

Eder, M., Rybicki, J. and Kestemont, M. (2015). 'Stylo': a package for stylometric analyses. file://prs-store2.unv.wlv.ac.uk/home2$/in4326/home/Profile/Downloads/stylo_howto%20(1).pdf . Accessed 17.5.17.

Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *R Journal* 8(1): 107-121.

Harris, M. (2010). Corpus Linguistics for the Decipherment of Rongorongo. *M. Res. Dissertation*, Birkbeck University.

Holmes, D. I., Gordon, L. J. and Wilson, C. (2001). A widow and her soldier: Stylometry and the American Civil War. *Literary and Linguistic Computing* 16(4): 403-420.

Hooper, W. (1971). *Past Watchful Dragons*. New York: Collier Books.

Hooper, W. (1977). Preface to *The Dark Tower and Other Stories* by C.S. Lewis. London: Fount.

Faust Jones, C. (1989). The Literary Detective computer analysis of stylistic differences between "The Dark Tower" and C. S. Lewis' Deep Space Trilogy. *Mythlore* **57**: 11-15

Juola, P. (2015).The Rowling case: A proposed standard analytic protocol for authorship questions. *Digital Scholarship in the Humanities*. **30, Supplement 1**: i100 to i113.

Lindskoog, K. (1978). Some problems in C.S. Lewis scholarship. *Christianity and Literature* 27(4), Summer 1978.

Lindskoog, K. (1988). *The C.S.Lewis Hoax*. Portland, Oregon: Multnomah Books.

Lindskoog, K. (1994). *Light in the Shadowlands*. Sisters, Oregon: Multnomah Books.

Lindskoog, K. (1999). Katherine Lindskoog's Informal Answer to the Ninth Non-Proof. The Lewis Legacy Issue 81, Summer 1999. The C.S. Lewis Foundation for Truth in Publishing. June 1, 1999. http://www.discovery.org/a/944. Accessed 16.5.17.

Linmans, A. J. M. (1998). Correspondence Analysis of the Synoptic Gospels. *Literary and Linguistic Computing*, 13(1):1-13.

Merriam, T. (2000). Edward III. Literary and Linguistic Computing, 15(2): 157-186.

Schöch, C. Principal Component Analysis for Literary Genre Stylistics, *The Dragonfly's Gaze*, September 29, 2013. http://dragonfly.hypotheses.org/472 . Accessed 29.6.17.

Tankard, J. (1986). The Literary Detective, *Byte*, February 1986: 231-238

Thisted, R. and Efron, B. (1987). Did Shakespeare write a newly discovered poem? *Biometrika* **74**(3):445-455.

Thompson, J. and Rasp, J. (2009). Did C. S, Lewis write The Dark Tower?: An examination 0f the small-sample properties of the Thisted-Efron tests of authorship. *Austrian Journal of Statistics*, **38**(2) : 71-82.

Ward, J. H., Jr. (1963). "Hierarchical Grouping to Optimize an Objective Function", Journal of the American Statistical Association, 58: 236–244.